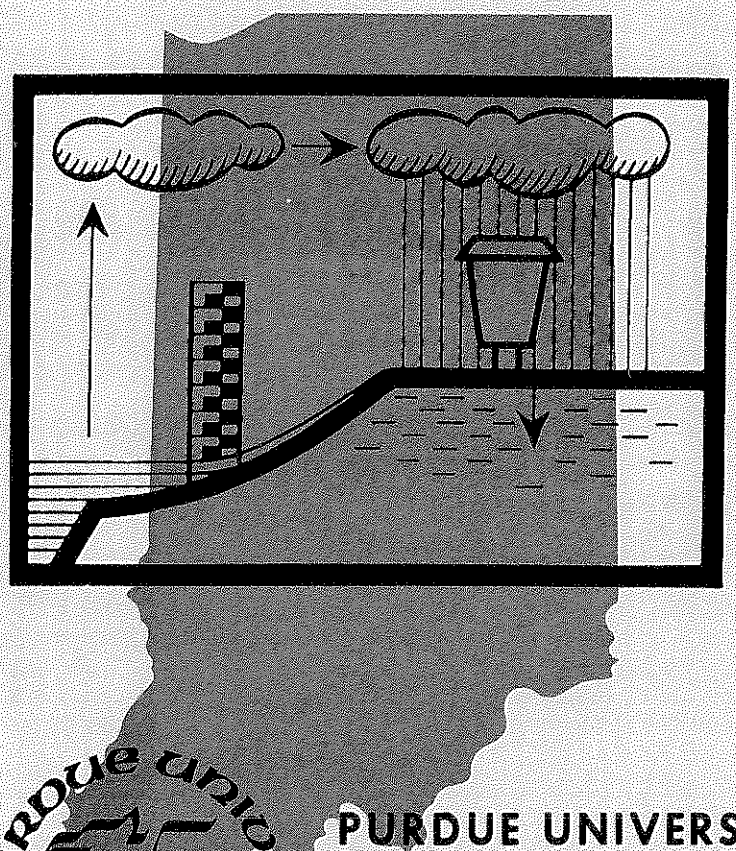
TECHNICAL REPORT NO. 70

AN INFORMATION RETRIEVAL SYSTEM FOR THE MACROINVERTEBRATE FAUNA OF INDIANA RIVERS AND LAKES



by

W. P. McCafferty

September 1975

PURDUE UNIVERSITY
WATER RESOURCES RESEARCH CENTER
WEST LAFAYETTE, INDIANA

WATER RESOURCES RESEARCH CENTER
PURDUE UNIVERSITY
West Lafayette, Indiana 47907

AN INFORMATION RETRIEVAL SYSTEM FOR THE MACROINVERTEBRATE FAUNA OF INDIANA RIVERS AND LAKES

by

W. P. McCafferty

The work upon which this report is based was supported in part by funds provided by the United States Department of the Interior, Office of Water Research and Technology, as authorized by the Water Resources Research Act of 1964 (PL 88-379 as amended).

Period of Investigation: January 1, 1975 - June 30, 1975
Completion Report for Project No. OWRT-A-041-IND

PURDUE UNIVERSITY WATER RESOURCES RESEARCH CENTER
TECHNICAL REPORT NO. 70

ABSTRACT

Title: An Information Retrieval System for the Macroinvertebrate Fauna of Indiana Rivers and Lakes. Dr. W. P. McCafferty, Principal Investigator

Spatial, temporal, and micro-habitat data are associated with comprehensive benthic and terrestrial samples of Ephemeroptera from throughout Indiana and taken over the past three and one-half years. Objectives of the research were to implement a utilitarian index system to provide the needed capability of correlating available subsets of environmental data. This model would then be available for making biological predictions concerning environmental water quality. The system is unique since it utilizes the natural ecological distribution of an entire aquatic fauna for a broad geographic area. A storage and retrieval system was designed after the Broadhurst Uniterm System. Various modifications of this system were made in order to accommodate the ecological and taxonomic data for which it has not been used previously. This system essentially consists of (1) the samples; (2) the accession index; and (3) the uniterm retrieval index. All information is completely cross referenced in the system. The system is open ended so that any new categories or information can be incorporated at any time. The model initially consists of 1,571 separate Indiana accessions and over 880 separate data uniterms within 17 categories and with an average of 91 references per uniterm for a total of over 80,000 accession numbers or reference inputs.

Descriptors: *Systematics, *Baseline Studies, Data Storage and Retrieval

Identifiers: *Ephemeroptera, *Informational Retrieval, Broadhurst Uniterm System, Retrieval Index, Retrieval System

AN INFORMATION RETRIEVAL SYSTEM FOR THE MACROINVERTEBRATE FAUNA OF INDIANA RIVERS AND LAKES

W. P. McCafferty

its immediate usage for not only general regional assessments of environmental quality but also the evaluation of more precise points throughout Indiana. The coordinate indexing system was designed as follows:

1. - Biological materials and associated data: Each vial of stored specimens represents a unique combination of a single species taken at a particular time, at a particular place, and with certain ecological conditions associated with it. Data labels remain with the specimens at all times. This automatically precludes the usage of any materials that have not been thoroughly sorted, identified, and verified by an authority. Each vial therefore represents a unique combination of data. These vials are curated and catalogued taxonomically via family, genus, and species. Each vial has also been assigned a consecutive accession number which remains with it and actually ties the biological materials directly to the second major component of the system, the accession index.

2. - The accession index. This is a separate serial card file or master file in which all the data associated with any one unique sample whether it be taxonomic or ecological is tabulated on a single card. This index is set up, however, irrespective of any classification or relationship of the data represented, but merely according to a serial number assigned in sequence to the unique samples as they are accessioned into the system.

The data from each card in the accession index and thus each unique data sample is "traced", that is, appropriate terms or phrases which delimit a particular parameter or range of parameter are singled out. For example, the particular species name, the temperature, the current discharge, the week of the year, the substrate type, etc. These key words or phrases are called "uniterms" and form the basis for the third integral component of the system, the uniterm retrieval index.

3. - The uniterm retrieval index: Uniterms which have been traced from the data sample are used as individual headings to cards or sets of cards which hence form the retrieval index. Under each appropriate heading the accession numbers of each unique sample which may apply are "posted". Proceeding in this way generated a complete system of cross indexed data. To facilitate

matching and correlation data sets, a card design for the uniterm cards was incorporated as follows: The cards consist of ten columns, numbered consecutively. Accession numbers are located in the columns not according to the first digit but the last. This enables the quick comparison of two or more cards by comparing appropriate columns. One provision of this system, however, is that numbers always be listed sequentially from smallest to largest in all of the columns. Therefore posting must be done in order of accession if interline insertions are to be avoided. Since the number of uniterms is already known for any sample, there should never be any need to retrace a sample for additional uniterms. An exception to this would be if the taxonomic nomenclature of a species changed with time. In this case the species name appearing on the accession card could simply be changed. If the name is synonymized, it would be best to void the previous accession numbers and cards which applied and simply reaccession the samples. This would also apply if information concerning the sample improved with time, e.g., concerning an ability to make more precise determinations.

The system at this time consists of over 1,800 accessions from approximately 11,000 specimens, and over 880 data uniterms have been utilized. There is an initial average of 95 references per uniterm for some 83,600 reference inputs. The uniterms can be divided into categories and subcategories (some of which are obviously uniterms within themselves) as follows:

A. Geographic
   1. State of Country
   2. County or Similar Unit
   3. Exact Locality
B. Temporal
   1. Year
   2. Week or Year
C. Taxonomic
   1. Family
   2. Genus
   3. Species

D. Biological
   1. Sex
   2. Growth Stage

E. Ecological
   1. Water System
   2. Water Classification
   3. Current Range
   4. Substrate Type
   5. Temperatures
   6. Nocturnal Light Attraction

F. Precision of Data
   1. Collectors
   2. Identifiers
   3. Reared
   4. Stages Associated
   5. Identified to Species

Consideration of any uniterm card provides the accession numbers of samples applicable to the parameter expressed by the uniterm itself. For two or more uniterms, appropriate uniterm cards are simply compared for coincidences of accession numbers. Making a series of such cross matches of uniterm cards yields information in as broad or fine a detail as is required.

The efficiency of the system is a function of several things. It is obviously directly related to the amount and comprehensiveness of the data incorporated. It is also limited by the complexity of manipulation which is expected to increase as data input increases, and it is obvious that a point will be reached at which it may become practical to employ computer methods to make searches. Over and above this, however, it should be pointed out that the efficiency of the system is also a function of the questions asked of it. The number of questions which could be asked of the system are inexaustible. Three examples where one can envision an application of the system to water quality anaysis are as follows: 1) Which species of Ephemeroptera have been taken

in the years 1971-1975 in a one-half mile stretch of Wildcat Creek which will represent the immediate effluent area of an impending impoundment? The pre- and post-impoundment analysis would be greatly aided by having baseline data already established and retrievable. 2) Which species of Ephemeroptera could be expected as mature larvae in early June in littoral regions of small lakes and ponds in LaGrange, Steuben, Noble and DeKalb Counties? If such species were to then be found to be conspicuously absent in certain of the lakes of this region and season, this empirical relationship may justify further environmental analysis of cause and effect. 3) In what streams is species X found whose range of tolerance for dissolved oxygen or some other factor is known to be quite low? Some indication of the water quality of certain streams could be biologically assessed in part by this data.

The immediate goals of developing a coordinate indexing system for freshwater macroinvertebrates has been met. The system is entirely applicable to any group of organisms for any area, and can incorporate any types of environmental data. The beauty of the system is that it is entirely "open ended", that is new categories, uniterms, and data can continue to be added at any time.

Long term objectives which have been beyond the scope of the project as funded are: 1. To test the utility of the system in predicting environmental quality of Indiana rivers and lakes, 2. To expand the data available for input, and 3. To completely automate the system via computer facilities available at Purdue University.